Prediction of radionuclide diffusion enabled by missing data imputation and ensemble machine learning*

Jun-Lei Tian,[1] Jia-Xing Feng,[1] Jia-Cong Shen,[1] Lei Yao,[1] Jing-Yan Wang,[1] Tao Wu,[1,†] and Yao-Lin Zhao[2,‡]

[1]*Huzhou Key Laboratory of Environmental Functional Materials and Pollution Control, Huzhou University, Huzhou 313000 China*
[2]*School of Nuclear Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China*

Missing values in radionuclide diffusion datasets can undermine the predictive accuracy and robustness of machine learning (ML) models. A regression-based missing data imputation method using light gradient boosting machine (LGBM) algorithm was employed to impute over 60% of the missing data, establishing a radionuclide diffusion dataset containing 16 input features and 813 instances. The effective diffusion coefficient(D_e) was predicted using ten ML models. The predictive accuracy of ensemble meta-models, namely LGBM-extreme gradient boosting (XGB) and LGBM-categorical boosting (CatB), surpassed the other ML models, with R^2 values of 0.94. The models were applied in predicting the D_e values of $EuEDTA^-$ and $HCrO_4^-$ in saturated compacted bentonites at compaction ranged from 1200 kg/m^3 to 1800 kg/m^3, which was measured using a through-diffusion method. The generalization ability of LGBM-XGB model surpassed that of LGB-CatB in predicting the D_e of $HCrO_4^-$. Shapley additive explanations identified the total porosity as the most significant influencing factor. In addition, the partial dependence plot analysis technique showed clearer results for univariate correlation analysis. This study provides a regression imputation technique to refine radionuclide diffusion datasets, offering a deeper insight into analyzing the diffusion mechanism of radionuclide and supporting the safety assessment of the geological disposal of high-level radioactive waste.

Keywords: machine learning; radionuclide diffusion; bentonite; regression imputation; missing data; diffusion experiments.

I. INTRODUCTION

Bentonite is often selected as an engineering barrier in a high-level radioactive waste (HLW) repository due to the low hydraulic conductivity, leading to a diffusion-controlled process for the transport of radionuclides [1–4]. The effective diffusion coefficient (D_e), a critical parameter in the safety assessment of repositories, describes the diffusion behavior of radionuclide in porous media [5–7]. Under complex disposal conditions, D_e is affected by the properties of radionuclides, such as diffusing species, adsorption properties [8], the characteristics of bentonite, such as compaction, pore structure, physical and chemical properties [3, 9, 10], and the porewater chemistry, such as pH and ionic strength [11–14]. Over the past decades, much attention has been devoted to determining the D_e of radionuclides in compacted bentonite [1, 8, 15–17].

Predicting the D_e of radionuclides is both challenging and crucial due to the nonlinear and complex interactions among radionuclides, porewater, and bentonite[2, 3]. Machine learning (ML) models are valuable tools for this task because they can manage complex and high-dimensional data. Various ML models, such as light gradient boosting machine (LGBM), extreme gradient boosting (XGB), categorical boosting (CatB), support vector machine (SVM), random forest (RF), and artificial neural networks (ANN), have been applied in predicting the D_e of radionuclides in compacted bentonite [18–21]. The radionuclide diffusion datasets were compiled from experimental data published in literatures and from a radionuclide diffusion database established by the Japan Atomic Energy Agency (JAEA-DDB). These datasets included the number of input features ranged from 3 to 16 and the data size ranged from 293 instances to 956 instances [19–21]. It is worth mentioning that the JAEA-DDB collected over 5000 instances from radionuclide diffusion experiments, spanned from 1982 to 2009 [22]. However, the instances increased with decreasing input features, primarily due to the missing data, resulting in potential impact on the accuracy and reliability of ML model explanations.

The issues caused by the presence of missing data are a pervasive concern in databases [23, 24]. Missing data can lead to suboptimal outcomes, reduce predictive performance, and even result in misleading conclusions [25, 26]. For instances, the dry density and rock capacity factor have been reported as the top-two influencing factors in predicting the D_e [20, 21]. In contrast, Wu et al. (2024) observed that the ion diffusion coefficient in water and dry density were observed as the top-two contributors. This discrepancy can be attributed to the insufficient number of instances in the datasets used. Therefore, a comprehensive dataset is essential for providing a more reliable analysis of the diffusion mechanism.

This study presents a novel, comprehensive radionuclide diffusion dataset with micro-mesoscopic features using ML models as regression imputation techniques. Firstly, LGBM was employed as a regression-based missing data imputation method to impute over 60% missing data. Subsequently, ten ML models, including three ensemble ML algorithms (LGBM-CatB, LGBM-XGB, and LGBM-RF), four decision tree algorithms (LGBM, CatB, XGB, and RF), Support Vector Machine (SVM), and two neural networks (ANN and deep neural network (DNN)), were trained, optimized, and tested using five-fold cross validation to predict D_e values. Finally, through-diffusion experiments were conducted to measure the diffusion parameters of $EuEDTA^-$ and $HCrO_4^-$ in

* This work was partially supported by the National Natural Science Foundation of China (No. 12475340 and 12375350), Special Branch project of South Taihu Lake, and the Scientific Research Fund of Zhejiang Provincial Education Department (No. Y202456326) .
† Corresponding author, Tao Wu: twu@zjhu.edu.cn
‡ Corresponding author, Yao-Lin Zhao: zhaoyaolin@mail.xjtu.edu.cn

compacted bentonite, including D_e, rock capacity factor, accessible porosity, total porosity, and distribution coefficient, to evaluate the generalization of the trained ML models. The goal is to develop predictive models that exhibit high accuracy, strong robustness, and clear interpretability for radionuclide diffusion studies, which are crucial for the safety assessment of HLW repositories.

II. MATERIALS AND METHODS

A. Material

Ba-bentonite was prepared by modifying Gaomiaozi (GMZ) bentonite with $BaCl_2$ solution. The mass percentage of $BaCl_2$ in modified bentonite was 5%. The detailed procedures for this modification are described in a previous study [16]. Wyoming bentonite powder had the grain dry density of 2760 kg/m^3, montmorillonite content of 0.85, external surface area of 38 m^2/g, and cation exchange capacity of 78.7 meq/100g [27, 28]. Ba-bentonite powder had the grain dry density of 2710 kg/m^3, montmorillonite content of 0.78, external surface area of 27.3 m^2/g, and cation exchange capacity of 58.7 meq/100g [16].

All solid chemicals were purchased from Aladdin. The pH values of NaCl solution were adjusted to 5.0 $\pm$ 0.1 and 7.0 $\pm$ 0.1 for $EuEDTA^-$ and $HCrO_4^-$ diffusion experiments, respectively. A stock solution of $EuEDTA^-$ was prepared by dissolving a measured amount of $EuNO_3{\cdot}6H_2O$ in 200 mL of a solution mixed with 0.6 mol/L NaCl and 0.01 mol/L EDTA. Similarly, a stock solution of $HCrO_4^-$ was prepared by dissolving a measured amount of $K_2Cr_2O_7$ in 200 mL of 0.5 mol/L NaCl solution. The initial concentrations of $HCrO_4^-$ and $EuEDTA^-$ were 1.8 $\times 10^{-3}$ mol/L and 5.7 $\times 10^{-4}$ mol/L, respectively, with corresponding pH values of 5.3 $\pm$ 0.1 and 6.8 $\pm$ 0.1. The uncertainty in pH was determined based on the standard deviation derived from five source solutions for $HCrO_4^-$ and $EuEDTA^-$. The excess EDTA ensured the complete complexation of Eu(III).

B. Through-diffusion method

A through-diffusion method was conducted to measure diffusion parameters of $EuEDTA^-$ and $HCrO_4^-$ in compacted bentonites. The experiments were operated under ambient conditions, with pH 5.3 $\pm$ 0.1 and a temperature of 25 $\pm$ 3°C for $EuEDTA^-$ diffusion, and pH 6.8 $\pm$ 0.1 and a temperature of 15 $\pm$ 3°C for $HCrO_4^-$ diffusion. Bentonite powder was compacted into cylindrical blocks with dry densities in the range of 1200$-$1800 kg/m^3. The powder, with an initial water content of approximately 5%, was calculated to weigh between 7.8 g and 11.4 g for the preparation of the bentonite blocks. During the weighing process and the preparation of bentonite blocks in the experimental procedure, approximately 0.3 g of bentonite powder was lost. This loss represents the primary source of uncertainty in the compacted dry density. Table 1 summarizes the experimental conditions for these diffusion experiments. After the compacted bentonite blocks were mounted in diffusion setups, they were saturated for five weeks with NaCl solution in diffusion cells. The diffusion experiments of lasted 90 days for $EuEDTA^-$ and 25 days for $HCrO_4^-$.

Table 1. Overview of the experimental condition for $EuEDTA^-$ and $HCrO_4^-$ diffusion experiments.

Experimental conditions	Detailed information	
Anion	$EuEDTA^-$	$HCrO_4^-$
Bentonite type	Ba-bent.	Wyoming
Initial concentration ($\times 10^{-3}$ mol/L)	0.57 $\pm$ 0.02	1.80 $\pm$ 0.10
Ionic strength (mol/L)	0.6	0.5
Dry density (kg/m^3)	1300$-$1700	1200$-$1800
pH ($-$)	5.3 $\pm$ 0.1	6.8 $\pm$ 0.1
Temperature (°C)	25 $\pm$ 3	15 $\pm$ 3
Block dimension (cm)	Ø 2.54 $\times$1.3	Ø 2.54 $\times$1.2
Volume of source reservoir (mL)	200	
Volume of target reservoir (mL)	10	

Concentrations of Cr and Eu were measured using an inductively coupled plasma optical emission spectrometer (Optima 7000DV, PerkinElmer, USA). Data processing was performed using Fitting for diffusion parameters software to calculate diffusion parameters, such as the D_e, rock capacity factor, distribution coefficient, total porosity, and accessible porosity. Further details regarding the experimental setup, operation steps, and data processing are available in previous studies[17, 29].

C. Data

1. *Data compilation*

Datasets were gathered from JAEA-DDB and 16 published resources, covering the period from 1982 to 2024. The dataset comprised 16 input features and 324 experimental instances, including 304 instances obtained from Wu et al. (2024) and an additional 20 experimental instances from three other literatures [17, 20, 27]. Notably, the absence of pH values in 514 instances of the JAEA-DDB resulted in a significantly reduction in data size. To address this, regression imputation techniques using ML models were applied to predict pH values based on the dataset of 324 instances, thereby expanding the dataset to 838 instances.

The dataset included 16 input features, which were categorized into three groups: (i) porewater properties, comprising the ionic strength (I), temperature (T), and pH; (ii) bentonite properties, including the montmorillonite content (m), external surface area (A_{ext}), dry density (ρ_d), grain density (ρ_s), total porosity (ε_{tot}), and montmorillonite stacking number (n_c); and (iii) radionuclide properties, encompassing the ion diffusion coefficient in water (D_w), molecular weight (MW), ion molar conductivity (λ), ionic radius (r), ionic charge (z), distribution coefficient (K_d), and rock capacity factor (α).

Table 2. Details of the features and instances of datasets.

Dataset	Input feature	Input number	Output feature	Instance number
Dataset I	Basic features: (i) Porewater: I, T. (ii) Bentonite: m, A_{ext}, ρ_d, ρ_s, ε_{tot}, n_c. (iii) Radionuclides: D_w, r, z, λ, MW, K_d, α.	15	pH	316
Dataset II	Basic features and pH	16	D_e	316
Dataset III	Basic features and pH	16	D_e	813

2. *Data preprocessing*

The presence of outliers can reduce predictive accuracy of ML models. To address this issue, the Mahalanobis distance (MD) method was employed to identify and remove outliers. The cutoff point (d_i) is given as:

$$d_i = \sqrt{(x-\mu)\cdot S^{-1}\cdot(x-\mu)}, \quad (1)$$

where x represents the object vector, μ denotes the mean arithmetic vector, and S is the covariance matrix of instances. The cutoff point was set to eight to ensure that the skewness of all input features was less than 10.

Three datasets were utilized to enhance the prediction of radionuclide diffusion. An overview of the features and instances for each dataset is summarized in Table 2. Dataset I included 15 input features, with pH as the output feature. To ensure data quality and reduce noise,eight instances were removed using MD method. This process yielded Dataset I, which comprised 316 instances. Statistical details for Dataset I are presented in Table S1 of the supporting information. Both Datasets II and III comprised 16 input features, including the basic features (15 input features of Dataset I) and pH. The output feature for Datasets II and III was the D_e. Dataset III, comprising 813 instances, was obtained by removal of 17 instances. It is noteworthy that these datasets comprised parameters at the micro-mesoscopic level. Specifically, the montmorillonite stacking number and ionic radius were classified as microscopic parameters, while other parameters were considered as mesoscopic.

3. *Imputation methods*

Four decision tree models, namely LGBM, CatB, XGB, and RF, were used as regression imputation methods to predict the pH values of Dataset I. LGBM exhibited superior predictive accuracy compared to the other models. This is consistent with our previous work [21]. Dataset III was established by incorporating additional 514 instances with Dataset II using LGBM for data imputation. Table S2 of the supporting information summarizes the statistical results of input and output features for Dataset III.

D. Methodology

The D_e of radionuclide in compacted bentonite was predicted using ten ML models, including three ensemble ML algorithms (LGBM-CatB, LGBM-XGB, and LGBM-RF), four decision tree algorithms (LGBM, CatB, XGB, and RF), SVM, and two neural networks (ANN and DNN). Ensemble ML models combine the strengths of multiple individual models to enhance overall predictive performance and stability, offering a promising solution to the challenges of bias and variance in individual models [30]. Since LGBM exhibited superior predictive performance compared to other models, it was employed to combine with CatB, XGB, and RF to predict the D_e using a voting regressor method from the scikit-learn package [20, 31]. The voting regressor simultaneously applies multiple regression models to the same dataset, thereby optimizing the final output by synthesizing the prediction results of each model. During the training process, the system can adjust the weight distribution according to the performance of each model. The final prediction result $\hat{y}$ is calculated by:

$$\hat{y} = \sum_{i=1}^{n} y_i \omega_i, \quad (2)$$

where y_i and ω_i represent the prediction result and the weight corresponding of the i-th model, respectively. This method optimized the weight ranges of base learners within a model by initially pruning these ranges according to the gradient of the best base learners performance, thereby accelerating the process of model optimization [30]. The hyperparameters of ML models were tuned using Particle Swarm Optimization (PSO) algorithm. In this algorithm, the potential solutions to an optimization problem are represented as a swarm of particles. Each particle i possesses a position vector X_i and a velocity vector V_i within the search space. During the algorithmic evolution, iterative adjustments are performed on both the velocity and position of each particle. Specifically, the velocity of each particle is updated according to the individual's best-known position p_i and the swarm's global best position g_i, as follows:

$$x_i^{k+1}(t+1) = x_i^k(t) + v_i^{k+1}(t+1), \quad (3)$$

$$\begin{aligned} v_i^{k+1}(t+1) = {} & \omega v_i^k(t) + c_1 r_1 \left(p_i^k(t) - x_i^\kappa(t)\right) \\ & + c_2 r_2 \left(g^k(t) - x_i^k(t)\right), \end{aligned} \quad (4)$$

where ω is inertia weight, which influences the particle's velocity based on its previous state. c_1 and c_2 represent the

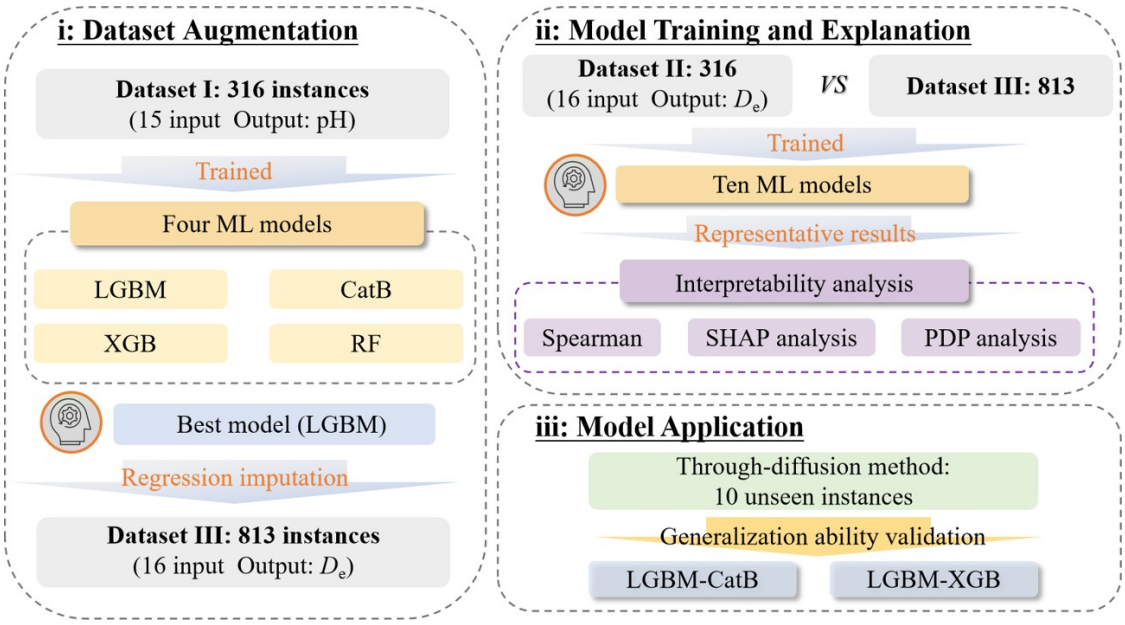

Fig. 1. Workflow diagram on building machine learning models for predicting the effective diffusion coefficient of radionuclides in various compacted bentonites.

learning factor for individual and social adjustment, respectively. r_1 and r_2 denote random numbers uniformly distributed within [0, 1].

Fig.1 illustrates a workflow diagram for developing ML models to predict the D_e values of radionuclides in various compacted bentonites. This work was organized into three parts: (i) Dataset augmentation: Missing pH values was predicted using decision tree algorithms, thereby refining the radionuclide diffusion dataset. (ii) Model training and explanation: Ten ML models were employed to train prediction models with high predictive accuracy. The diffusion mechanism was analyzed using Spearman, Shapley additive explanations (SHAP), and partial dependence plots (PDP). (iii) Model application: The D_e of $EuEDTA^-$ and $HCrO_4^-$ in compacted bentonites was measured using a through-diffusion method, which was employed to evaluate the generalization capability of the best ML models.

E. Model development and evaluation

Datasets were randomly divided into a training set consisting 80% of the instances and a test set containing the remaining 20%. Since the data processing using logarithmic transformation and min-max normalization exhibited insignificant impact on the predictive accuracy in predicting the D_e of radionuclides in bentonite [19], logarithmic transformation was applied to the features, such as the ionic radius, ion diffusion coefficient in water, and D_e, due to their significantly larger magnitudes compared to other features. A five-fold cross validation method was used to decrease the risk of overfitting. Therefore, the 80% training data was further subdivided into a pretraining dataset (80% of the training data) and a validation dataset (20% of the remaining training data) to pretrain ML models and optimize hyperparameters. The PSO technique was conducted to facilitate the optimization of hyperparameters.

The predictive performance was evaluated by the coefficient of determination (R^2), and mean square error (MSE). These metrics are given as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\log D_{e,i}^{\exp} - \log D_{e,i}^{\mathrm{pred}}\right)^2}{\sum_{i=1}^{N}\left(\log D_{e,i}^{\exp} - \log D_{e,ave}^{\exp}\right)^2}, \tag{5}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\log D_{e,i}^{\exp} - \log D_{e,i}^{\mathrm{pred}}\right)^2, \tag{6}$$

where $\log D_{e,i}{}^{\exp}$ and $\log D_{e,ave}{}^{\exp}$ are the experimental D_e and average experimental D_e measured from diffusion experiments. $\log D_{e,i}{}^{\mathrm{pred}}$ is the predicted D_e using the ML models.

III. RESULTS AND DISCUSSION

A. Model development

1. *Regression imputation for predicting pH*

Handling missing data is a crucial step that can significantly impact the quality and reliability of data analysis. Various regression imputation techniques have been applied

Table 3. Mean performance metric values using five-fold cross validation and the highest performance metrics for machine learning models to predict pH based on Dataset I.

Algorithms	Datasets	R^2_{cv}	MSE_{cv}	Best performance R^2	Best performance MSE
LGBM	Training	**0.99**	0.01	**0.99**	0.01
	Validation	**0.87**	0.32	**0.90**	0.07
	Test	**0.88**	0.33	**0.92**	0.23
XGB	Training	0.98	0.05	0.98	0.06
	Validation	0.82	0.46	0.92	0.16
	Test	0.84	0.47	0.87	0.38
CatB	Training	0.99	0.01	0.99	0.01
	Validation	0.87	0.28	0.86	0.22
	Test	0.83	0.68	0.85	0.57
RF	Training	0.90	0.27	0.90	0.26
	Validation	0.77	0.61	0.79	0.67
	Test	0.77	0.38	0.80	0.32

for imputing missing data, such as ANNs, multivariate imputation by chained equations, k-nearest neighbors, time-series deep learning model, generative broad Bayesian imputer, principal component analysis imputation, and simple arithmetic averages. These methods have been applied to datasets with missing data percentages ranging from 0 to 80% [24, 26, 32–36]. Generally, three types of missing data mechanisms are recognized, namely missing completely at random, missing at random, and missing not at random [23]. Each mechanism presents different challenges and implications for the imputation, highlighting the importance of identifying the underlying pattern of missingness before selecting an appropriate imputation strategy.

JAEA-DDB database collected the data from literatures and reports, covering the period from 1982 to 2009. The instances are derived from various diffusion experimental methods and numerous researchers. The absence of pH values in 514 instances within the JAEA-DDB database can be explained that researches ignored the importance of pH values in their studies. In the JAEA-DDB database, missing data primarily resulted from ignoring or inadequately measuring the parameters that related to the radionuclide diffusion. The missing mechanism in the JAEA-DDB database was assumed to be the missing completely at random, corresponding to noncontinuous missingness. Based on the selected 16 input features, more than 60% of the dataset (514 instances) lack pH values. Decision tree models were employed to predict the missing pH values, aiming to augment the dataset and enhance the robustness of ML models. Specifically, LGBM, CatB, XGB, and RF were employed to predict pH values for Dataset I.

The predicted performance is summarized in Table 3. LGBM exhibited superior robustness compared to the other models. For instances, the R^2_{cv} values for the test sets were ranked in descending order using five-fold cross validation as follows: LGBM > XGB > CatB > RF. The rank of MSE_{cv} values were in the opposite with R^2_{cv} values for the test datasets. Notably, LGBM achieved the highest performance metrics among all models, with a MSE of 0.23 and R^2 of 0.92 for the test dataset, respectively. The hyperparameters of the best ML models are listed in Table S3 of the supporting information. Therefore, the missing pH values for 514 instances were predicted using the LGBM model, resulting in the establishment of Dataset III with 813 instances.

Fig.2 exhibited data distribution and characteristics of the relationship between pH and each input feature. Blue and orange represent the data distribution of Dataset I and imputed 514 instances, respectively. It clearly demonstrates that there is non-linear relationship between pH and each input feature. The predicted pH values ranged from 5.0 to 9.0, exhibiting a Gaussian type distribution.

pH is an important porewater parameter that influences both radionuclide species and the surface charge of clay [37]. Fig.3 shows the dependency of pH on the external surface area and ion molar conductivity, which are associated with bentonite and radionuclide properties, respectively. Dataset I exhibits that the pH value ranged from 3.0 to 13.4. The predicted pH values were concentrated in the range from 5.0 to 9.0, suggesting a close adherence to a normal distribution of porewater for Dataset III.

2. *Model development for radionuclide diffusion*

Ten ML models, namely LGBM-CatB, LGBM-XGB, LGBM-RF, LGBM, CatB, XGB, RF, ANN, DNN, and SVM, were conducted for predicting the D_e of radionuclide in compacted bentonite. Fig.4 shows the performance metrics of the ML models for test datasets of Dataset II and III, using the optimal hyperparameters tuned with PSO techniques (Table S4 in the supporting information). The performance metrics were assessed using five-fold cross validation. The red lines represent the kernel smooth curve of the distribution of performance metrics. The black lines within and outside the box plots denote the mean values and standard deviation of the performance metrics, respectively, with a lower standard deviation indicating strong robustness of ML models. Detailed performance metrics for training datasets, validation datasets, and test datasets can be found in Table S5 of the supporting information.

As the number of instances increased from 316 (Dataset II) to 813 (Dataset III), the performance metrics of all ML models improved significantly, as evidenced by higher R^2_{cv} values, lower MSE_{cv}, and reduced standard deviation. These findings indicate that expanding the dataset contributed to enhanced predictive performance and robustness of the ML models. It is noteworthy that the ensemble models were established by combining LightGBM with other individual decision tree models, primarily due to the relatively high training speed of the LightGBM algorithm [38]. However, there was no significant difference in computational efficiency between the ensemble model and the single model. The difference in running time was approximately five minutes. In the case of decision tree algorithms, gradient boosting (GB) models (LGBM, CatB, and XGB) outperformed the RF models. The excellent predictive performance of GB models is consistent with previous findings in predicting chloride diffusion

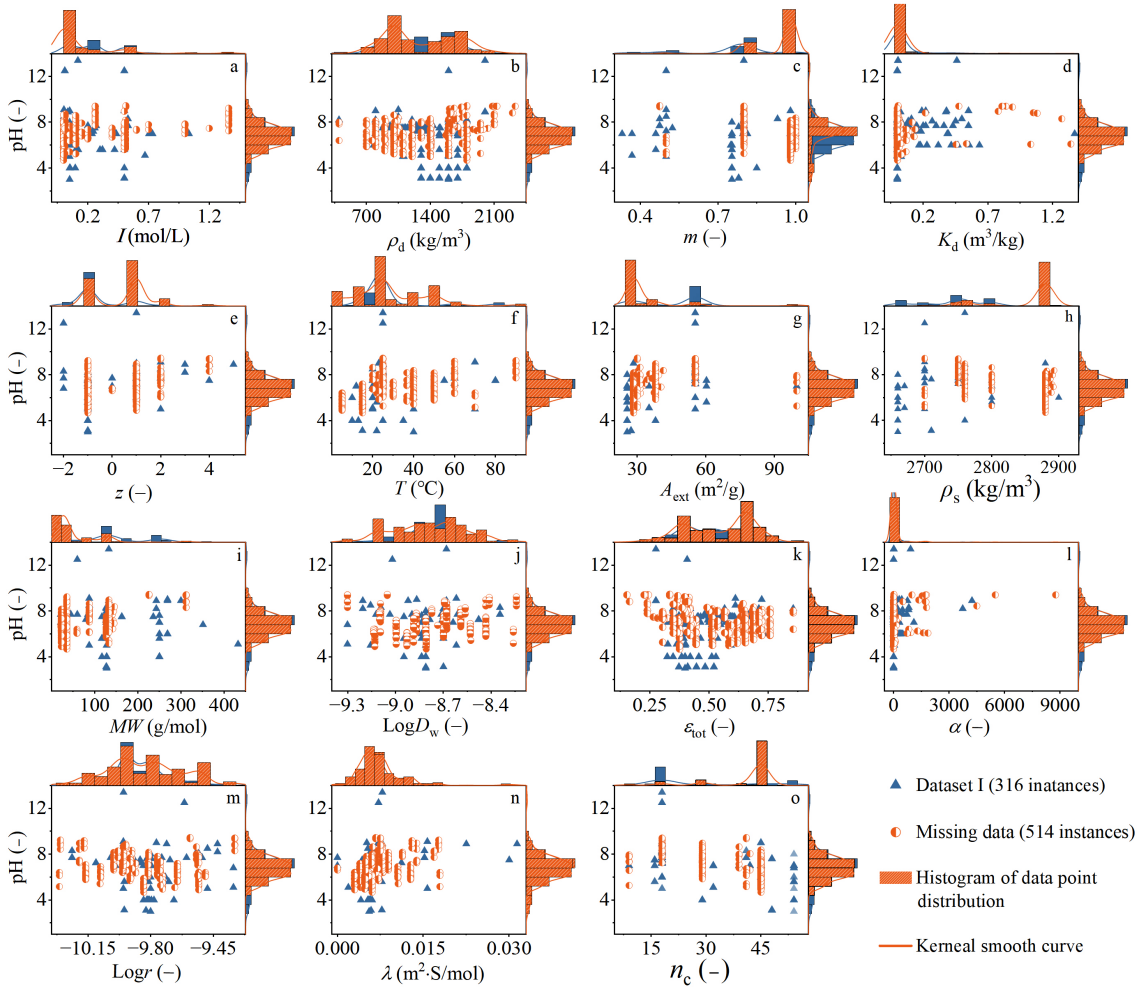

Fig. 2. Data distribution of features and the relationship between pH and each input feature.

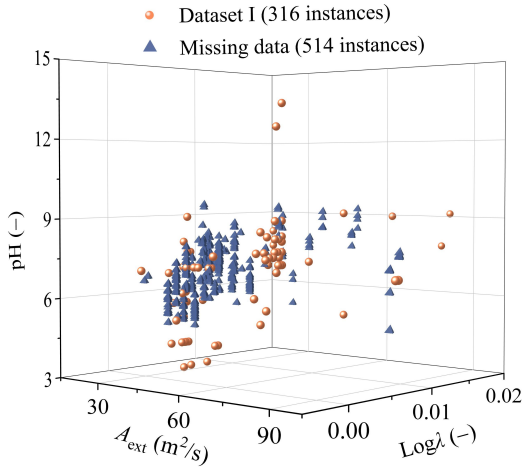

Fig. 3. Analyzing the dependency of pH on the external surface area and ion molar conductivity.

coefficient in concrete [39]. In addition, ensemble ML models (LGBM-CatB, LGBM-XGB, and LGBM-RF) and LGBM surpassed the other ML models, achieving R^2_{cv} above 0.90. It can be attributed to their capability of harnessing the strengths of various diverse algorithms to thoroughly capture potentially complex patterns and errors within the data, thereby enhancing prediction accuracy and robustness [30]. For Dataset III, the R^2_{cv} values of the ML models ranked in descending order as follows: LGBM-CatB ≈ LGBM-XGB > LGBM ≈ LGBM-RF > CatB ≈ XGB > ANN > DNN > RF > SVM. Notably, LGBM-CatB surpassed LGBM-XGB due to its lower standard deviation, indicating stronger robustness. SVM exhibited the lowest predictive performance based on Dataset III, with R^2_{cv} = 0.75 and MSE_{cv} = 0.06. Compared with the ensemble models, the SVM is a relatively simple model. These ensemble models are designed to capture more complex patterns and relationships in the data through the combination of multiple decision trees. This lack of complexity in SVM limits its ability to generalize well across different data instances in the dataset. Notably, some studies have reported the test R^2 values below 0.80, such as an R^2 of 0.74 for predicting the retention rate of Cd in biochar [40] and an R^2 of 0.76 for predicting alcohol space-time yield [41]. Therefore, the prediction accuracy of SVM remained satisfactory,

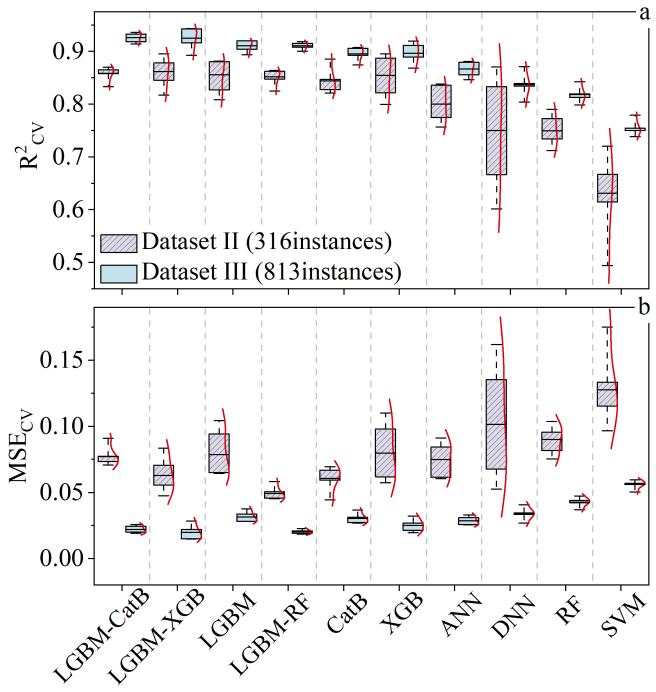

Fig. 4. Mean performance metric values using five-fold cross validation for machine learning models in the test datasets of Dataset II and III.

despite exhibiting lower predictive performance compared to other models.

Fig.5 shows the regression plots comparing the experimental and predicted D_e values for the training (green triangle), validation (red circle), and test (purple square) datasets of Datasets II and III, using LGBM-CatB, LGBM-XGB, LGBM, and LGBM-RF algorithms. These algorithms were selected for their excellent predictive accuracy. The plots reveal a close alignment between the experimental and predicted D_e values with the slope line, underscoring the effective simulation capability of these ML models in predicting radionuclide diffusion processes. The performance metrics for the best-performing models are also displayed in Fig.5. Notably, the ML models applied to the test dataset of Dataset III outperformed those applied to Dataset II. This disparity can likely be attributed to the augmentation of instances in Dataset III, which facilitated the models' ability to more effectively capture complex relationships within the data. For Dataset III, the ranking of models was as follows: LGBM-CatB (R^2 = 0.94) ≈ LGBM-XGB (R^2 = 0.94) > LGBM (R^2 = 0.92) ≈ LGBM-RF (R^2 = 0.92). These results indicate that both LGBM-CatB and LGBM-XGB exhibited high predictive accuracy.

B. Sensitivity analysis

1. *Spearman and Shapley additive explanation analyses*

ML models can uncover predictive principles through analysis techniques that rank the importance of influencing factors on predictions, such as feature importance and SHAP analysis [19, 21, 42, 43]. Additionally, Spearman analysis, a non-parametric statistical method, assesses the monotonic relationship between two variables by correlating ranked data. These approaches provided valuable insights into the consistency and strength of relationships within the dataset. It worthy notes that the reliability of these analytical techniques is intrinsically linked to the quality of the data used. Increasing the dataset size enhances the depth, broadness, and reliability of ML models.

Spearman correlation and SHAP analyses techniques were employed to analyze the correlation and importance of input features, presented intuitively global interpretations of ML models (Fig.6). The features are ranked from left to right according to their correlation and the contribution to the prediction. The Spearman correlation analysis reveals that the most influencing factor among the 16 input features was the ion diffusion coefficient in water for Dataset II and total porosity for Dataset III . This feature exhibited a positive correlation with D_e (Figs.6a and b). This is consistent with the previous finding [19] and Archie's law[31, 44].

In the case of Dataset II, the SHAP analysis reveals that the most important input feature varied across different ML models: the compacted dry density for LGBM-CatB, ionic radius for LGBM-XGB, and ion diffusion coefficient in water for LGBM (Figs.6c, e, and g). Notably, the SHAP results for LGBM were the only ones consistent with the Spearman correlation analysis. This discrepancy can be attributed to the differences in feature importance assessment and prediction mechanisms inherent to each ML algorithm. As the number of instances increased from 316 (Dataset II) to 813 (Dataset III), both Spearman and SHAP analyses identified the total porosity as the primary contributor, which is consistent with Archie's law[31, 44]. The total porosity for radionuclide diffusion in compacted bentonite blocks is expressed as a percentage of the total interconnected pore spaces within the blocks. A higher total porosity implies a greater availability of transport pathways. These findings suggest that larger datasets may reduce the discrepancies between ML models in terms of feature importance assessment and prediction mechanisms.

2. *Partial dependence plots*

The dependency of D_e on the 16 input features has been discussed in our previous work [19]. However, some relationships may remain unclear due to the limited size of dataset. To address this, a PDP analysis was performed to visually represent the univariate correlations and to examine how the size of the dataset influences these relationships (Fig.7). The histograms and lines correspond to the data distribution and the correlation with each input feature and the PDP. A more concentrated data distribution generally leads to more accu-

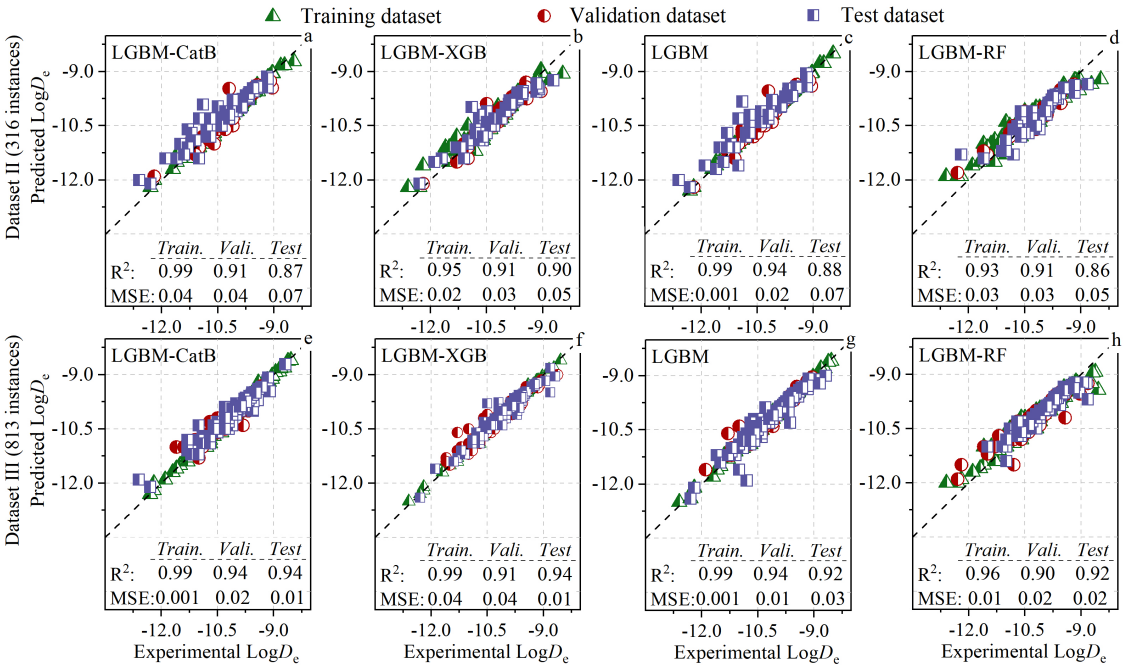

Fig. 5. Regression plots of experimental versus predicted effective diffusion coefficients based on Datasets II and III: (a, e) LGBM-CatB, (b, f) LGBM-XGB, (c, g) LGBM, and (d, h) LGBM-RF.

rate analysis results. These findings indicate that Dataset III, which is larger than Dataset II, exhibits more continuous PDP curves, suggesting a more stable and clear relationship between the features and D_e.

Figs.7a and b shows that both rock capacity factor and distribution coefficient exhibited clear positive correlation with the prediction for Dataset III. This finding aligns with studies on radionuclides diffusion in crystalline rock [45] and sodium montmorillonite [46]. Consistently, Fig.7d illustrates a positive impact of ionic charge, where cations exhibit higher D_e than neutral species, and anions display lower D_e values. This is consistent with previous studies, which attributed the differences in diffusion mechanisms to electrostatic interactions between radionuclide species and charged bentonite surfaces [3]. Specifically, cation diffusion is controlled by surface diffusion effects, whereas anions diffusion is driven by anionic exclusion effects [46, 47].

pH values in the range from 6 to 9 exhibited a negative influence on the prediction for Dataset III, while a peak was observed at approximately pH 8 for Dataset II (Fig.7c). The negative impact of Dataset III might be more convincing due to the larger data size. Fig.7e shows a positive impact on the prediction when ion molar conductivity exceeded 0.01 m^2·S/mol for Dataset III. However, the relationships of external surface area, montmorillonite stacking number, grain density, and ionic strength remained unclear for both Dataset II and III (Figs.7f−i). This lack of clarity can be attributed to the dispersion of data, despite larger dataset size.

In the case of remaining input features, such as the total porosity, ion diffusion coefficient in water, and temperature, exhibited positive impacts on the prediction, whereas the dry density, montmorillonite content, ionic radius, and molecular weight showed negative impacts (Figs.7j−p). The positive influence of the total porosity and ion diffusion coefficient in water could be explained by Archie's law [16, 44], while the positive impact of temperature followed Arrhenius equations [48–50]. The detailed explanations can be found in our previous studies [19, 21]. It is worth mentioning that a negative influence of ionic radius was observed at Log$r$ < −9.6 (2.5 Å). The positive relationship can be attributed to the limited data for species with ionic radius above 2.5 Å. Overall, the univariate correlation results visualized using the PDP technique align with the diffusion laws observed in experiments and the diffusion mechanisms derived from numerical models. This consistency underscores the reliability of the interpretation capabilities of the ML models.

C. Diffusion experiments and model application

Anionic radionuclides with long half-life are very important in the safety evaluation of HLW repository because of their high diffusivity. A through-diffusion method was employed to measure the diffusion parameters of $EuEDTA^-$ and $HCrO_4^-$ in compacted bentonites at compacted dry density ranged from 1200 kg/m^3 to 1800 kg/m^3. Their D_e values were predicted using LGBM-CatB and LGBM-XGB to test the generalization ability.

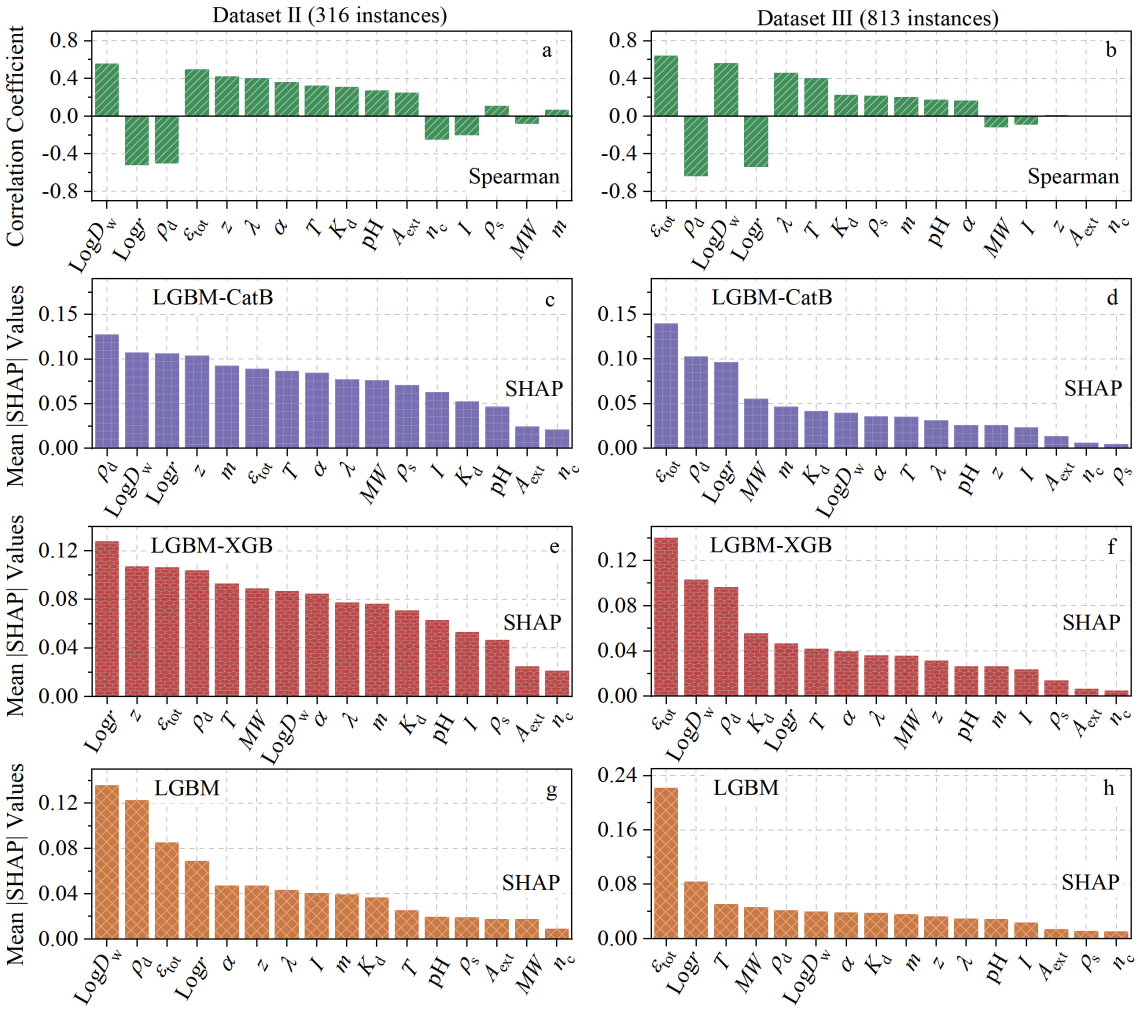

Fig. 6. (a, b) Spearman correlation analysis and global interpretations of ML models based on Dataset II and III: (c, d) LGBM-CatB, (e, f) LGBM-XGB, and (g, h) LGBM.

1. *Determination of the diffusion parameters using diffusion experiments*

Fig.8 shows the breakthrough curves of $EuEDTA^-$and the species distribution of Eu-EDTA complexes. A_{cum} denotes the accumulated mass of $EuEDTA^-$ and $HCrO_4^-$ that penetrated a 1.2 cm thick bentonite block to reach sample reservoirs. The data show that the accumulated mass increased with decreasing dry density, consistent with the general understanding that lower dry density facilitates radionuclide diffusion through porous media [3, 5]. The pH was maintained at 5.3 $\pm$ 0.1 during the Eu(III) diffusion experiments. Simulation using Vision MINTEQ indicated that Eu(III) exists as a mixture of species in 0.6 mol/L NaCl solution, including Eu^{3+}, EuHEDTA(aq), $EuEDTA^-$, and $EuCl^{2+}$ (Fig.8c). $EuEDTA^-$ was the main species at pH above 2.0. It indicates that this study measured the diffusion parameters of $EuEDTA^-$ in compacted Ba-bentonite.

Table 4 summarizes the diffusion parameters of $HCrO_4^-$ and $EuEDTA^-$, including D_e, rock capacity factor, accessible porosity, total porosity, and distribution coefficient. Both D_e and distribution coefficient are two important parameters in the safety assessment of repositories, while the other parameters play a crucial role in elucidating the diffusion mechanism. The error in the compacted dry density measurement was primarily attributed to a loss of approximately 0.3 g during the preparation of bentonite blocks. Both $HCrO_4^-$ and $EuEDTA^-$ are monovalent anions that are unable to access the interlayer pores of compacted bentonite[17, 21]. The rock capacity factor of $HCrO_4^-$ was found to be lower than the total porosity, indicating that the accessible porosity was equaled to the rock capacity factor. This suggests that the predominant diffusion path of $HCrO_4^-$ was the free pores of compacted bentonite. In contrast, $EuEDTA^-$ exhibited adsorptive behavior similar to that of simulated trivalent actinide complexes, such as $AmEDTA^-$ and $CmEDTA^-$, with the rock capacity factor being higher than the total porosity. The distribution coefficient, K_d, of $EuEDTA^-$ was calculated as follows:

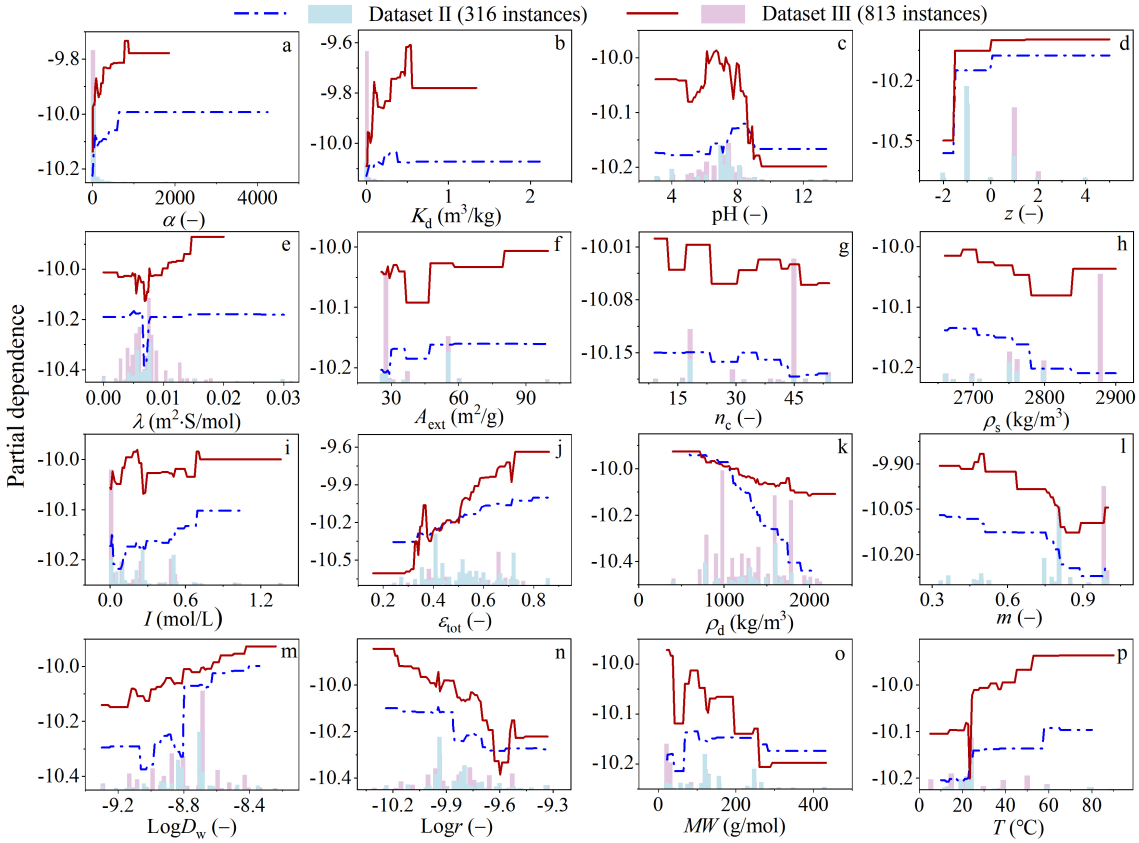

Fig. 7. Partial dependence plot for (a) the rock capacity factor, (b) distribution coefficient, (c) pH, (d) ionic charge, (e) ion molar conductivity, (f) external surface area, (g) montmorillonite stacking number, (h) grain density, (i) ionic strength, (j) total porosity, (k) dry density, (l) montmorillonite content, (m) ion diffusion coefficient in water, (n) ionic radius, (o) molecular weight, and (p) temperature.

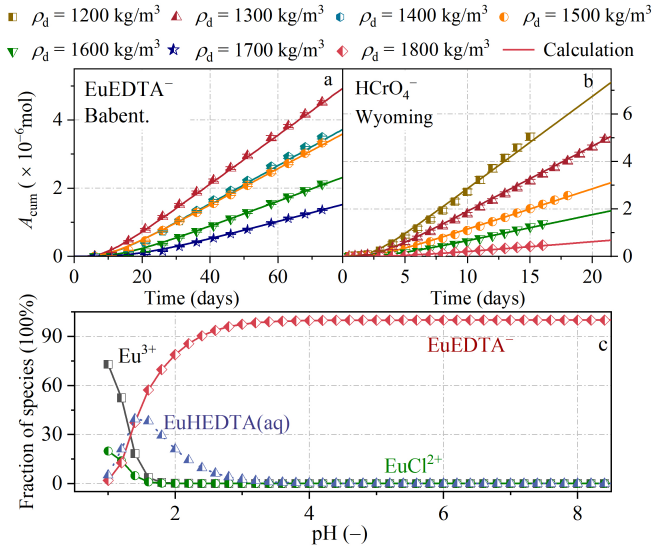

Fig. 8. The relationship between the accumulated mass (A_{cum}) and time for (a) $EuEDTA^-$ and (b) $HCrO_4^-$ in saturated compacted bentonites. (c) Species distribution of Eu(III)-EDTA system in aqueous solution.

$$K_d = \frac{\alpha - \varepsilon_{acc}}{\rho_d}, \tag{7}$$

where the accessible porosity, ε_{acc}, was obtained using the I^- diffusion experiments [19].

All the diffusion parameters decreased with increasing dry density for both $EuEDTA^-$ and $HCrO_4^-$. The distribution coefficient of $EuEDTA^-$ ranged from 4.2×10^{-4} m^3/kg to 6.7×10^{-4} m^3/kg, which is lower than the range reported for $EuEDTA^-$ in hard rock clay (1.3×10^{-3}–3.2×10^{-3} m^3/kg) [51] and for $CeEDTA^-$ in compacted Zhisin bentonite (0.8×10^{-3}–1.2×10^{-3} m^3/kg)[17] . The distribution coefficient of $EuEDTA^-$ was less than Eu^{3+}, indicating that EDTA facilitated the diffusion of Eu(III), thereby reducing the retardation capacity of the bentonite barrier[51, 52]. This observation is consistent with the diffusion behaviors of $CeEDTA^-$ and $CoEDTA^{2-}$ [17, 19, 31].

2. *Model application*

The LGBM-CatB and LGBM-XGB models were employed to predict the D_e of $HCrO_4^-$ in compacted Wyoming bentonite and $EuEDTA^-$ in compacted Ba-bentonite, which were compared with published diffusion experimental re-

Table 4. Overview of diffusion parameters of $EuEDTA^-$ and $HCrO_4^-$ in compacted bentonite.

ρ_d (kg/m^3)	m_{bent} (g)	D_e ($\times10^{-11}$ m^2/s)	D_a ($\times10^{-11}$ m^2/s)	α (−)	ε_{acc} (−)	ε_{tot} (−)	K_d ($\times10^{-4}$m^3/kg)
$EuEDTA^-$ in Ba-bentonite							
1300 ± 45	8.7 ± 0.3	3.6 ± 0.4	3.0 ± 0.3	1.2 ± 0.1	0.33 ± 0.01[#]	0.52	6.7 ± 0.6
1400 ± 45	9.3 ± 0.3	2.8 ± 0.3	2.6 ± 0.2	1.1 ± 0.1	0.31 ± 0.01[#]	0.48	5.6 ± 0.6
1500 ± 46	9.8 ± 0.3	2.6 ± 0.3	2.7 ± 0.2	1.0 ± 0.1	0.30 ± 0.01[#]	0.45	4.7 ± 0.5
1600 ± 46	10.5 ± 0.3	1.8 ± 0.2	1.9 ± 0.1	1.0 ± 0.1	0.26 ± 0.01[#]	0.41	4.3 ± 0.5
1700 ± 47	11.2 ± 0.3	1.3 ± 0.1	1.5 ± 0.1	0.9 ± 0.1	0.19 ± 0.01[#]	0.37	4.2 ± 0.3
$HCrO_4^-$ in Wyoming bentonite							
1200 ± 46	7.8 ± 0.3	6.2 ± 0.6	11.9 ± 0.5	0.52 ± 0.04	0.52 ± 0.04	0.57	–
1300 ± 52	7.7 ± 0.3	3.9 ± 0.3	8.1 ± 0.3	0.48 ± 0.04	0.48 ± 0.04	0.53	–
1500 ± 45	10.0 ± 0.3	2.7 ± 0.2	10.2 ± 0.2	0.26 ± 0.02	0.26 ± 0.02	0.46	–
1600 ± 47	10.2 ± 0.3	1.8 ± 0.1	7.7 ± 0.2	0.23 ± 0.02	0.23 ± 0.02	0.42	–
1800 ± 47	11.4 ± 0.3	0.7 ± 0.1	5.7 ± 0.1	0.12 ± 0.01	0.12 ± 0.01	0.35	–

[#] Data from [19]

sults for $HCrO_4^-$ and simulated actinides $CeEDTA^-$ and $CoEDTA^{2-}$[17, 19, 21]. Additionally, both models were conducted in predicting the D_e of radionuclide cation $^{137}Cs^+$ and neutral species HTO [8, 53, 54] (Fig.9). It shows that D_e/D_w decreased with increasing compacted dry density, which is consistent with previous studies[3, 5, 44]. In this study, the D_w value for metal-EDTA complexes was assumed to be 5.0 $\times$ 10^{-10} m^2/s[55]. The D_e of $EuEDTA^-$ was observed to be higher than that of $CeEDTA^-$ [17]. and $CoEDTA^{2-}$[19]. The LGBM-CatB and LGBM-XGB models demonstrated successful prediction of D_e, as evidence by the good agreement with the experimental D_e values (Fig.9a).

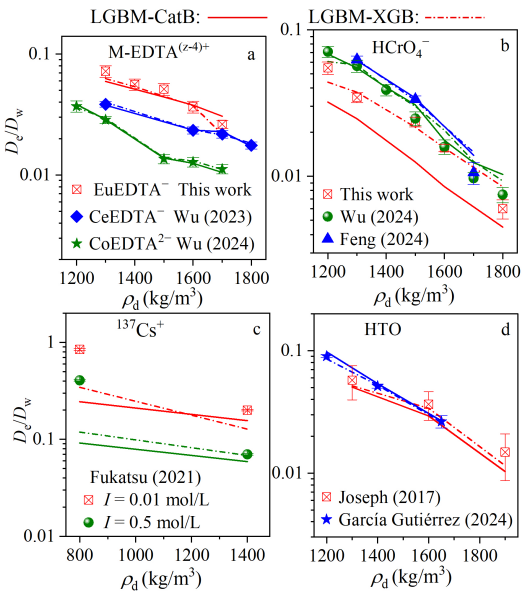

Fig. 9. Generalization ability validation of LGBM-CatB and LGBM-XGB: (a) M-EDTA$^{(z-4)+}$ diffusion, (b) $HCrO_4^-$ diffusion, (c) $^{137}Cs^+$ diffusion, and (d) HTO diffusion. .

Fig.9b shows that the D_e of $HCrO_4^-$ in compacted Wyoming bentonite was found to be lower than that in Anji bentonite [19] and GMZ bentonite[21] , likely due to the higher montmorillonite content. LGBM-CatB slightly underestimated D_e for $HCrO_4^-$ in Wyoming bentonite, with predicted D_e values being 25%−47% lower than the experimental D_e. Although this discrepancy is less pronounced compared to the predictions for $HCrO_4^-$ in GMZ and Anji bentonites using LGBM and PSO-LGBM, where the difference was reported to be 9%−27%[19, 21]. This performance is significantly superior to the prediction of Archie's law, which reported that the predictive D_e values were 1.0 to 1.5 orders of magnitude higher than experimental results[44].

Fig.9c shows that the predicted D_e values of $^{137}Cs^+$ are consistent with the experimental results at a compacted density of 1400 kg/m^3. However, a significant underestimation was observed at the compacted density of 800 kg/m^3, with the difference being approximately four times. It can be explained by the limited number of experimental data points available for this density in the dataset, which comprised only 58 instances, accounting for approximately 7% of the total dataset. It indicates that more diffusion experiments for $^{137}Cs^+$ should be conducted at the compacted density around 800 kg/m^3 to facilitate the identification of diffusion patterns by ML models. Fig.9d illustrates that both LGBM-CatB and LGBM-XGB models exhibit accurate prediction for the D_e of HTO. Under similar experimental conditions, the D_e in Wyoming bentonite (red squares) was higher than that in FEBEX bentonite (blue pentagrams), primarily attributed to the lower montmorillonite content, with $m = 0.85$ for Wyoming bentonite and $m = 0.92$ for FEBEX bentonite [53, 54].

Notably, the experimental diffusion data from this study, as well as from $^{137}Cs^+$ [8] and HTO [53, 54] diffusions , were not included in the test datasets, which highlights the strong generalization ability of both LGBM-CatB and LGBM-XGB models. Furthermore, the generalization ability of LGBM-XGB was superior to that of LGBM-CatB, indicating that model selection plays a crucial role in accurately predicting radionuclide diffusion in complex geological environments. Given that HLW repositories are designed to operate for over 10,000 years, the prediction of radionuclide diffusion in bentonite barriers must consider the complex coupling effect

among radionuclides, porewater, and bentonite under intrinsic disposal conditions. Current diffusion datasets remain insufficient for the safety assessment of bentonite barriers due to limitations in data size and dimensionality. Therefore, more diffusion experiments should be conducted to enhance the dimensionality and scale of datasets.

IV. CONCLUSION

A radionuclide diffusion dataset, comprising 16 input features and 813 instances, was developed using regression imputation machine learning (ML) methods. Ten ML algorithms was employed to predict the effective diffusion coefficient (D_e) of radionuclides in compacted bentonite. The light gradient boosting machine (LGBM)-extreme gradient boosting (XGB) and LGBM-categorical boosting (CatB) algorithms surpassed the other ML models, achieving R^2 values of 0.94 based on the imputed dataset. This improvement indicates that the imputed dataset enabled the ML models to achieve high predictive performance and strong robustness.

The generalization of LGBM-CatB and LGBM-XGB models was evaluated by applying them to predict the D_e of $EuEDTA^-$ in compacted Ba-bentonite and $HCrO_4^-$ in compacted Wyoming bentonite. Both models exhibited excellent predictive accuracy of $EuEDTA^-$, while LGBM-CatB slightly underestimated D_e for $HCrO_4^-$ in Wyoming bentonite, with predicted D_e values being 25%−47% lower than the experimental D_e. This indicates that the generalization ability of LGBM-XGB surpassed that of LGBM-CatB.

It has been widely accepted that the quality and quantity of datasets plays a crucial role for the predictive performance of ML models. However, a significant number of diffusion experimental results have been excluded from diffusion datasets due to incomplete or missing data. To address this limitation, additional experiments are necessary to comprehensively characterize the properties of porewater and bentonite. These experiments should include, but are not limited to, mineral composition analysis, elemental analysis, and particle size analysis..

V. AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Jun-Lei Tian, Jia-Xing Feng and Yao-Lin Zhao. The first draft of the manuscript was written by Tao Wu and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

VI. BIBLIOGRAPHY

[1] L. Baborová, E. Viglaová, D. Vopálka, Cesium transport in Czech compacted bentonite: Planar source and through diffusion methods evaluated considering non-linearity of sorption isotherm. Appl. Clay Sci. **245**, 107150 (2023). https://doi.org/10.1016/j.clay.2023.107150

[2] L. Cui, W. Ye, Q. Wang et al., A model for describing advective and diffusive gas transport through initially saturated bentonite with consideration of temperature. Eng. Geol. **323**, 107215 (2023). https://doi.org/10.1016/j.enggeo.2023.107215

[3] P. Krejci, T. Gimmi, L. R. Van Loon et al., Relevance of diffuse-layer, Stern-layer and interlayers for diffusion in clays: A new model and its application to Na, Sr, and Cs data in bentonite. Appl. Clay Sci. **244**, 107086 (2023). https://doi.org/10.1016/j.clay.2023.107086

[4] R. Zuo, Z. Xu, X. Wang et al., Adsorption characteristics of strontium by bentonite colloids acting on claystone of candidate high-level radioactive waste geological disposal sites. Environ. Res. **213**, 113633 (2022). https://doi.org/10.1016/j.envres.2022.113633

[5] M. García Gutiérrez, J. Cormenzana, T. Missana et al., Diffusion coefficients and accessible porosity for HTO and ^{36}Cl in compacted FEBEX bentonite. Appl. Clay Sci. **26**, 65−73 (2004). https://doi.org/10.1016/j.clay.2003.09.012

[6] H. Lyu, Z. Xu, J. Zhong et al., Machine learning-driven prediction of phosphorus adsorption capacity of biochar: Insights for adsorbent design and process optimization. J. Environ. Manage. **369**, 122405 (2024). https://doi.org/10.1016/j.jenvman.2024.122405

[7] Y. Yang, S. V. Churakov, R. A. Patel et al., Pore-scale modeling of water and ion diffusion in partially saturated clays. Water Resour. Res. **60**, e2023WR035595 (2024). /https://doi.org/10.1029/2023WR035595

[8] Y. Fukatsu, K. Yotsuji, T. Ohkubo et al., Diffusion of tritiated water, $^{137}Cs^+$, and $^{125}I^-$ in compacted Ca-montmorillonite: Experimental and modeling approaches. Appl. Clay Sci. **211**, 106176 (2021). https://doi.org/10.1016/j.clay.2021.106176

[9] A. Asaad, F. Hubert, E. Ferrage et al., Role of interlayer porosity and particle organization in the diffusion of water in swelling clays. Appl. Clay Sci. **207**, 106089 (2021). https://doi.org/10.1016/j.clay.2021.106089

[10] C. Wigger, L. R. Van Loon, Importance of interlayer equivalent pores for anion diffusion in clay-rich sedimentary rocks. Environ. Sci. Technol. **51**, 1998−2006 (2017). https://doi.org/10.1021/acs.est.6b03781

[11] A. Muurinen, O. Karnland, J. Lehikoinen, Ion concentration caused by an external solution into the porewater of compacted bentonite. Phys. Chem. Earth. **29**, 119−127 (2004). https://doi.org/10.1016/j.pce.2003.11.004

[12] P. Wersin, M. Kiczka, K. Koskinen, Porewater chemistry in compacted bentonite: Application to the engineered buffer barrier at the Olkiluoto site. Appl. Geochem. **74**, 165−175 (2016). https://doi.org/10.1016/j.apgeochem.2016.09.010

[13] P. Wersin, M. Mazurek, T. Gimmi, Porewater chemistry of Opalinus clay revisited: Findings from 25 years of data collection at the Mont Terri Rock Laboratory. Appl. Geochem. **138**, 105234 (2022). https://doi.org/10.1016/j.ap-

geochem.2022.105234
[14] C. Wigger, L. R. Van Loon, Effect of the pore water composition on the diffusive anion transport in argillaceous, low permeability sedimentary rocks. J. Contam. Hydrol. **213**, 40−48 (2018). https://doi.org/10.1016/j.jconhyd.2018.05.001
[15] I. C. Bourg, A. C. Bourg, G. Sposito, Modeling diffusion and adsorption in compacted bentonite: a critical review. J. Contam. Hydrol. **61**, 293−302 (2003). https://doi.org/10.1016/S0169-7722(02)00128-6
[16] T. Wu, Z. Feng, Z. Geng et al., Restriction of Re(VII) and Se(IV) diffusion by barite precipitation in compacted bentonite. Appl. Clay Sci. **232**, 106803 (2023). https://doi.org/10.1016/j.clay.2022.106803
[17] T. Wu, Y. Hong, D. Shao et al., Experimental and modeling study of the diffusion path of Ce(III)-EDTA in compacted bentonite. Chem. Geol. **636**, 121639 (2023). https://doi.org/10.1016/j.chemgeo.2023.121639
[18] Z. Feng, Z. Gao, Y. Wang et al., Application of machine learning to study the effective diffusion coefficient of Re(VII) in compacted bentonite. Appl. Clay Sci. **243**, 107076 (2023). https://doi.org/10.1016/j.clay.2023.107076
[19] T. Wu, J. Tian, X. Shi et al., Predicting anion diffusion in bentonite using hybrid machine learning model and correlation of physical quantities. Sci. Total Environ. **946**, 174363 (2024). https://doi.org/10.1016/j.scitotenv.2024.174363
[20] X. Shi, J. Tian, J. Shen et al., Application of machine learning in predicting the apparent diffusion coefficient of Se(IV) in compacted bentonite. J. Radioanal. Nucl. Chem. **333**, 5811−5821 (2024). https://doi.org/10.1007/s10967-024-09637-w
[21] Z. Feng, J. Tian, T. Wu et al., Unveiling the Re, Cr, and I diffusion in saturated compacted bentonite using machine-learning methods. Nucl. Sci. Tech. **35**, 93 (2024). https://doi.org/10.1007/s41365-024-01456-8
[22] Y. Tochigi, Y. Tachi, Development of diffusion database of buffer materials and rocks-expansion and application method of foreign buffer materials. JAEA-Data/Code 2009−029. (2010). Japan Atomic Energy Agency,
[23] H. N. Haliduola, F. Bretz, U. Mansmann, Missing data imputation using utility-based regression and sampling approaches. Comput. Meth. Prog. Bio. **226**, 107172 (2022). https://doi.org/10.1016/j.cmpb.2022.107172
[24] W. S. Loh, L. Ling, R. J. Chin et al., A comparative analysis of missing data imputation techniques on sedimentation data. Ain Shams Eng. J. **15**, 102717 (2024). https://doi.org/10.1016/j.asej.2024.102717
[25] Y. Kim, S.M. Yi, J. Heo et al., Is replacing missing values of PM2. 5 constituents with estimates using machine learning better for source apportionment than exclusion or median replacement? Environ. Pollut. **354**, 124165 (2024). https://doi.org/10.1016/j.envpol.2024.124165
[26] M. Pastorini, R. Rodríguez, L. Etcheverry et al., Enhancing environmental data imputation: A physically-constrained machine learning framework. Sci. Total Environ. **926**, 171773 (2024). https://doi.org/10.1016/j.scitotenv.2024.171773
[27] Z. Feng, J. Tian, X. Shi et al., Analyzing porosity of compacted bentonite via through diffusion method. J. Radioanal. Nucl. Chem. **333**, 1185−1193 (2024). https://doi.org/10.1007/s10967-024-09368-y
[28] A. Idiart, M. Pkala Models for diffusion in compacted bentonite. SKB TR−16−06 (2016). Swedish Nuclear Fuel and Waste Management Company.
[29] T. Wu, Y. Yang, Z. Wang et al., Anion diffusion in compacted clays by pore-scale simulation and experiments. Water Resour. Res. **56**, e2019WR027037 (2020). https://doi.org/10.1029/2019WR027037
[30] N. Hou, Y. Tong, M. Zhou et al., New Strategies for constructing and analyzing semiconductor photosynthetic biohybrid systems based on ensemble machine learning models: Visualizing complex mechanisms and yield prediction. Bioresour. Technol. **412**, 131404 (2024). https://doi.org/10.1016/j.biortech.2024.131404
[31] Z. Feng, J. Feng, J. Tian et al., Predicting the diffusion of $CeEDTA^{-}$ and $CoEDTA^{2-}$ in bentonite using decision tree hybridized with particle swarm optimization algorithms. Appl. Clay Sci. **262**, 107596 (2024). https://doi.org/10.1016/j.clay.2024.107596
[32] S. C. Kuok, K. V. Yuen, T. Dodwell et al., Generative broad Bayesian (GBB) imputer for missing data imputation with uncertainty quantification. Knowl. Based Syst. **301**, 112272 (2024). https://doi.org/10.1016/j.knosys.2024.112272
[33] M. J. Kim, Y. Cho, Imputation of missing values in well log data using k-nearest neighbor collaborative filtering. Comput. Geosci. **193**, 105712 (2024). https://doi.org/10.1016/j.cageo.2024.105712
[34] J. C. Carpenter, Machine learning aids imputation of missing petrophysical data in Iraqi reservoir. J. Pet. Technol. **76**, 5861 (2024). https://doi.org/10.2118/0824-0058-JPT
[35] J. H. B. Abdulkhaleq, K. A. Khalil, W. J. Al Mudhafar et al., Advanced machine learning for missing petrophysical property imputation applied to improve the characterization of carbonate reservoirs. Geoemgry Sci. Eng. **238**, 212900 (2024). https://doi.org/10.1016/j.geoen.2024.212900
[36] G. Antariksa, R. Muammar, A. Nugraha et al., Deep sequence model-based approach to well log data imputation and petrophysical analysis: A case study on the West Natuna Basin, Indonesia. J. Appl. Geophy. **218**, 105213 (2023). https://doi.org/10.1016/j.jappgeo.2023.105213
[37] J. Yang, Z. Zhang, Z. Chen et al., Co-transport of U(VI) and gibbsite colloid in saturated granite particle column: role of pH, U (VI) concentration and humic acid. Sci. Total Environ. **688**, 450−461 (2019). https://doi.org/10.1016/j.scitotenv.2019.05.395
[38] Z. Gao, Y. Wang, H. Lü et al., Machine learning the nuclear mass. Nucl. Sci. Tech. **32**, 109 (2021). https://doi.org/10.1007/s41365-021-00956-1
[39] V. Q. Tran, Machine learning approach for investigating chloride diffusion coefficient of concrete containing supplementary cementitious materials. Constr. Build. Mater. **328**, 127103 (2022). https://doi.org/10.1016/j.conbuildmat.2022.127103
[40] J. Li, L. Pan, Z. Li et al., Unveiling the migration of Cr and Cd to biochar from pyrolysis of manure and sludge using machine learning. Sci. Total Environ. **885**, 163895 (2023). https://doi.org/10.1016/j.scitotenv.2023.163895
[41] M. Suvarna, P. Preikschas, J. Pérez Ramírez, Identifying descriptors for promoted rhodium-based catalysts for higher alcohol synthesis via machine learning. ACS catalysis. **12**, 15373−15385 (2022). https://doi.org/10.1021/acscatal.2c04349
[42] T. Liu, H. Zhang, J. Wu et al., Wastewater treatment process enhancement based on multi-objective optimization and interpretable machine learning. J. Environ. Manage. **364**, 121430 (2024). https://doi.org/10.1016/j.jenvman.2024.121430
[43] J. Zhang, Z. Long, Z. Ren et al., Application of machine learning in ultrasonic pretreatment of sewage sludge: Pre-

diction and optimization. Environ. Res. **263**, 120108 (2024). https://doi.org/10.1016/j.envres.2024.120108.

[44] L. R. Van Loon, J. Mibus, A modified version of Archie's law to estimate effective diffusion coefficients of radionuclides in argillaceous rocks and its application in safety analysis studies. Appl. Geochem. **59**, 85−94 (2015). https://doi.org/10.1016/j.apgeochem.2015.04.002

[45] H. Aromaa, M. Voutilainen, J. Ikonen et al., Through diffusion experiments to study the diffusion and sorption of HTO, ^{36}Cl, ^{133}Ba and ^{134}Cs in crystalline rock. J. Contam. Hydrol. **222**, 101−111 (2019). https://doi.org/10.1016/j.jconhyd.2019.03.002

[46] Y. Tachi, K. Yotsuji, Diffusion and sorption of Cs^+, Na^+, I^- and HTO in compacted sodium montmorillonite as a function of porewater salinity: Integrated sorption and diffusion model. Geochim. Cosmochim. Acta. **132**, 75−93 (2014). https://doi.org/10.1016/j.gca.2014.02.004

[47] M. Glaus, S. Frick, L. Van Loon, A coherent approach for cation surface diffusion in clay minerals and cation sorption models: Diffusion of Cs^+ and Eu^{3+} in compacted illite as case examples. Geochim. Cosmochim. Acta. **274**, 79−96 (2020). http://dx.doi.org/10.1016/j.gca.2020.01.054

[48] P. Chen, L. R. Van Loon, S. Koch et al., Reactive transport modeling of diffusive mobility and retention of TcO_4^- in Opalinus clay. Appl. Clay Sci. **251**, 107327 (2024). https://doi.org/10.1016/j.clay.2024.107327

[49] T. Kozaki, N. Saito, A. Fujishima et al., Activation energy for diffusion of chloride ions in compacted sodium montmorillonite. J. Contam. Hydrol. **35**, 67−75 (1998). https://doi.org/10.1016/S0169-7722(98)00116-8

[50] L. Van Loon, W. Müller, K. Iijima, Activation energies of the self-diffusion of HTO, $^{22}Na^+$ and $^{36}Cl^-$ in a highly compacted argillaceous rock (Opalinus clay). Appl. Geochem. **20**, 961−972 (2005). https://doi.org/10.1016/j.apgeochem.2004.10.007

[51] M. Descostes, I. Pointeau, J. Radwan et al., Adsorption and retarded diffusion of Eu^{III}-$EDTA^-$ through hard clay rock. J. Hydrol. **544**, 125−132 (2017). https://doi.org/10.1016/j.jhydrol.2016.11.014

[52] R. Dagnelie, P. Arnoux, J. Radwan et al., Perturbation induced by EDTA on HDO, Br^- and Eu^{III} diffusion in a large-scale clay rock sample. Appl. Clay Sci. **105**, 142−149 (2015). https://doi.org/10.1016/j.clay.2014.12.004

[53] M. García Gutiérrez, M. Mingarro, T. Missana, Influence of temperature and dry density coupled effects on HTO, ^{36}Cl, ^{85}Sr and ^{133}Ba diffusion through compacted bentonite. Prog. Nucl. Energy**. 176**, 105407 (2024). https://doi.org/10.1016/j.pnucene.2024.105407

[54] C. Joseph, J. Mibus, P. Trepte et al., Long-term diffusion of U(VI) in bentonite: Dependence on density. Sci. Total Environ. **575**, 207−218 (2017). https://doi.org/10.1016/j.scitotenv.2016.10.005

[55] K. Furukawa, Y. Takahashi, H. Sato, Effect of the formation of EDTA complexes on the diffusion of metal ions in water. Geochim. Cosmochim. Acta. **71**, 4416−4424 (2007). https://doi.org/10.1016/j.gca.2007.07.009